Exercise 4: Looking Critically at Your Dataset

Discussion Prompts

Use the prompts below to guide your group's conversation.
You can focus on a real research project or make one up for this exercise.

➔ **If you are an oversight committee member,** consider how you might use or adapt these questions in your review process—for example, by including them in application materials for researchers.

Getting Oriented

1. **Let's begin by getting clear on some high-level details to make the next questions easier.**

   On your reflection worksheet or in a separate document, jot down the following for each dataset you are using:

   a. Dataset name
   b. Number of people included
   c. Types of data collected
   d. Any obvious limitations or biases
   e. For virtual discussions: link to the dataset descriptor (if available)

   Consider arranging this information in a table, like the example below. This format can also be useful for sharing in a public forum:

Dataset name	**Number of people included at the time of your research**	**Types of data**	**Key limitations or biases**	**Link to dataset descriptor**
All of Us	800,000	Medical records, biosamples, genetic, wearable, omic	People living in US only	researchallofus.org/data-tools/data-snapshots/

Limitations of the Dataset

What data are available has an important impact on what research is undertaken.

2. How does the structure of the dataset(s) you will use in your research impact the focus of your research and/or the types of analysis you are able to do?

3. Have others raised concerns about this dataset? Do you have concerns about its integrity (for example, accuracy, completeness, or reliability)?

4. As you review the dataset's community labels or descriptors, can you imagine if and how any of them might be offensive to the communities they attempt to describe?

5. Are there known biases embedded in its collection (e.g., "race correction" algorithms, IQ measurements, or clinical tools that misperform across populations)?
    a. How will you account for them?

6. Describe any ongoing discourse or advocacy related to the categories used in this dataset (e.g., recent NASEM recommendations, OMB census revisions).

Consent and Participation

7. What metadata is available about the dataset(s) (e.g. data nutrition labels, datasheets)?

8. Can you access the dataset's informed consent document or a summary (e.g., in the documentation, metadata, or descriptors)?

9. Did participants consent on their own behalf, or was surrogate consent used (in the case of cognitively impaired participants and minors)?
    a. Or is there a mix of both self-consented and surrogate-consented participants?

10. Were participants compensated for providing their data? If so, how?

11. Does the dataset include requirements for monetary or non-monetary benefit-sharing back to participants or communities?
12. Was any data collected during patient care (e.g., EHR)?
    a. If yes, how might social dynamics have influenced what was recorded?

       For example, a trans patient may be seeking access to gender affirming care, a patient who uses drugs or alcohol may be embarrassed to report this to their provider, a minor patient may not report certain behaviors in the presence of their parents.

Transparency

13. Transparency is a core value of this work. We strongly encourage you to share your reflections and dataset descriptions in places where communities and other stakeholders can see them.

    This might include:

    a. A project website or institutional transparency page
    b. Community newsletters or plain-language project updates
    c. Social media

    Take a few minutes to note where and how you could share this information.

Further Reading (Optional)

If you'd like to explore further, here are some external resources.

This list includes resources embedded in the online version of this exercise. To access these, either search the titles below online or visit the web version for clickable links.

Dataset Biases & Categorization
Referenced in *Limitations of the Dataset*

- Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms
  (on "race correction" algorithms)

- The Eugenic Origins of IQ Testing: Implications for Post-Atkins Litigation
  (on IQ measurements)
- Racial Disparity in Oxygen Saturation Measurements by Pulse Oximetry: Evidence and Implications
  (on clinical tools that misperform across populations)

Population Descriptors & Standards
Referenced in *Limitations of the Dataset*

- Using Population Descriptors in Genetics and Genomics Research
  (on recent NASEM recommendations)
- Initial Proposals For Updating OMB's Race and Ethnicity Statistical Standards
  (on OMB census revisions)

Dataset Documentation & Transparency
Referenced in *Consent and Participation*

- The Data Nutrition Project
  (on data nutrition labels)
- Datasheets for Datasets
  (on datasheets)

Benefit-Sharing
More information on benefit-sharing

- Benefit-sharing - Nagoya Protocol HuB
  (on monetary or non-monetary benefit-sharing - Referenced in *Consent and Participation*)
- Benefit-Sharing by Design: A Call to Action for Human Genomics Research

Data Management Plans
Practical guides for developing data management plans

- Research Data Management - University of Cambridge
- Data Management - MIT
- Data Management - NIH